# Sharing Data from Human Subjects: Practical Realities

*Greg Farber*

*Office of Technology Development and Coordination, NIMH*

November 12, 2013

National Institute of Mental Health

# What is the Problem?

- Genomics studies have allowed great strides in uncovering the basis for a number of diseases that are caused by a mutation to a single gene.
- For complex diseases (multiple genes involved and may have an environmental component as well), neither genomics nor imaging (MRI) have been very successful at finding biomarkers that have a strong correlation to diagnosis or treatment.
- For many complex diseases a single diagnosis (diabetes, schizophrenia…) really covers a wide range of diseases that share some symptoms.

Potential Solutions:

- Continued experiments to uncover the basic biology.
- Combining data from experiments in multiple laboratories.

# Issues with Data from Human Subjects

- Experiments involving human subjects must conform to a number of ethical guidelines:
  - Belmont Report
  - Declaration of Helsinki
  - Council for International Organizations of Medical Sciences
- and regulations
  - HHS Regulations for the Protection of Human Subjects (45 CFR part 46)
  - Health Insurance Portability and Accountability Act (HIPAA) Regulations (45 CFR parts 160 and 164)
  - HHS Regulations for Responsibility of Applicants for Promoting Objectivity in Research for Which PHS Funding Is Sought (42 CFR part 50)
  - FDA Regulations for the Protection of Human Subjects (21 CFR part 50)
  - FDA Regulations for Institutional Review Boards (21 CFR part 56)

# More Issues with Data from Human Subjects

- Research laboratories often use multiple different clinical instruments to collect similar data.
- In many areas, a research subject may be seen in multiple laboratories, and unless those laboratories are willing to trade personally identifiable information about their research participants, it isn't possible to correctly combine datasets.

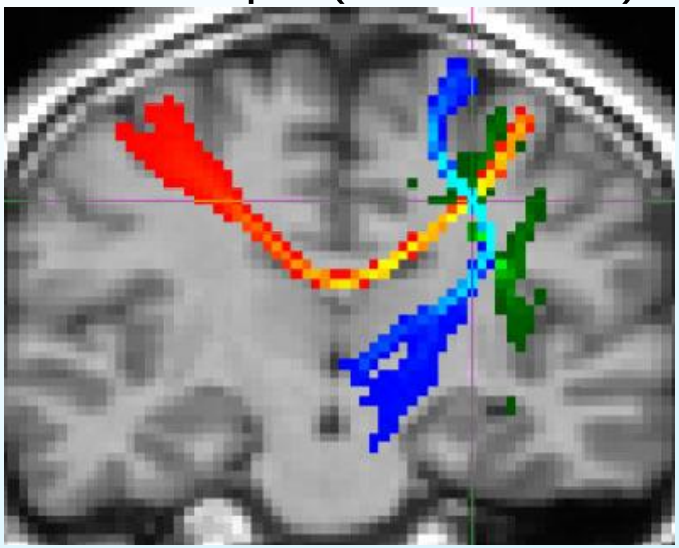
# Case Studies: Ways to Obey the Rules and Broadly Share Data from Human Subjects

- Human Connectome Project: behavioral, clinical (sensitive), genomic, and imaging data from a set of 1200 healthy young adults (identical twins and non-twin siblings).
  - Data collected using a single protocol
  - Data distributed from Washington University
  - Re-identification of subjects wouldn't be too hard – especially for participants!
- National Database for Autism Research: contains virtually all data from human subjects that is collected by grants funded by NIH
  - Many different data types are collected by different laboratories
  - Sharing PII between laboratories is a big problem
  - Other autism data repositories exist
  - Data distributed from a federal repository
  - Re-identification of subjects would be pretty difficult

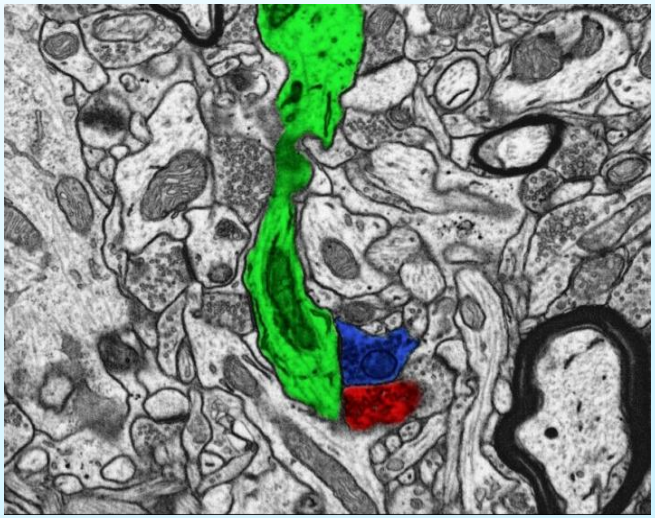
# Human Connectome Project: Basics

- A good definition of a connectome is a comprehensive map of neural connections.

Macroscopic (whole-brain)



Microscopic
(synapses between neurons)



The HCP is measuring the macro-connectome and its variability in healthy young adults.

# Human Connectome Project

- The Human Connectome is an NIH Blueprint funded project that is managed at NIMH.  There are two awards.  Bruce Rosen (MGH) is working on technology development.  David Van Essen (Wash U) also has a technology development component, but he is focused on collecting data.
- The Wash U – U Minnesota consortium is a true collaboration that includes investigators from 10 institutions.  Key contributions are coming from Oxford University, St. Louis University, and Indiana University.
- 99 on the HCP team.

# Human Connectome Project: Project Goals

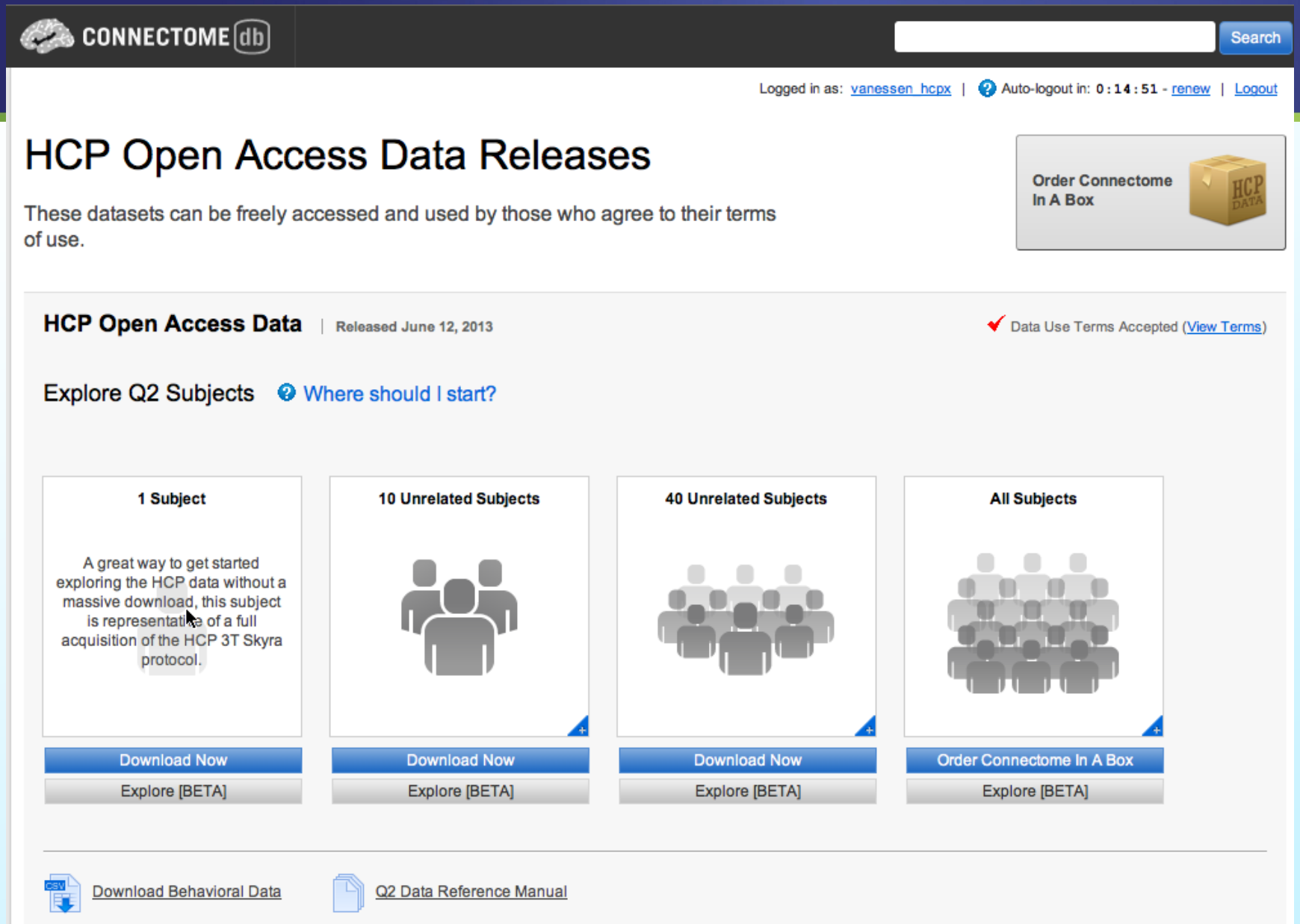
- Promote Data Mining, Discovery Science
  - Make data freely available
  - Implement a user-friendly informatics platform
- Relate brain connectivity to individual capabilities
  - Link to heritability, genetics
- Create a baseline for future studies of brain disorders
  - Autism, schizophrenia, ADHD, etc.

# Human Connectome Project: Implementation

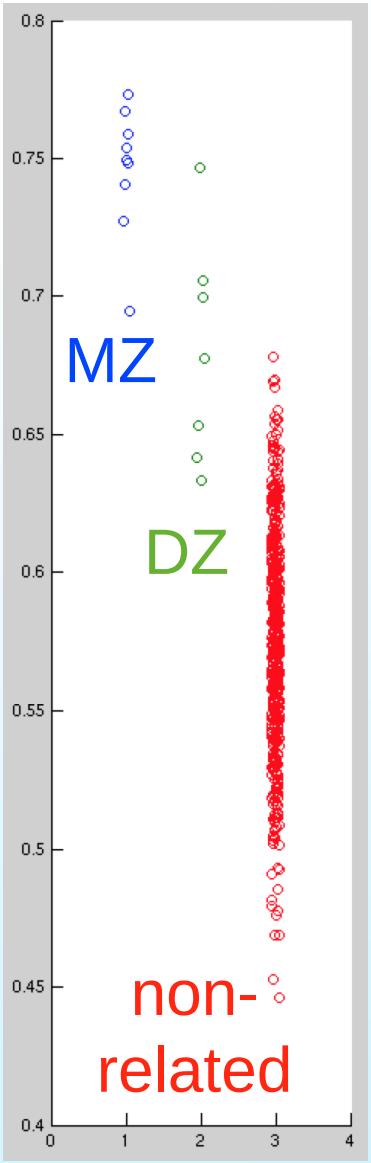
1. Very broad consent forms for participants.
2. Two tiered consent for researchers who want to use the data
   - Self registration and consent to “treat the data with respect” for non-sensitive data.
   - Basically anyone with a valid e-mail can get access to the non-sensitive data.
   - A second class of sensitive data has been established. This includes information that would make re-identification easy such as exact age, body weight, height, ethnicity, family structure. Information about psychiatric and neurological illness, drug use, and some other clinical data are also sensitive.
   - Sensitive data can be obtained by submitting a slightly more detailed request.
   - In all cases, those using the data are warned about obtaining appropriate institutional approval to use the data, but they do not need to submit certification that they have obtained those approvals.

# Human Connectome Data: Access to Data

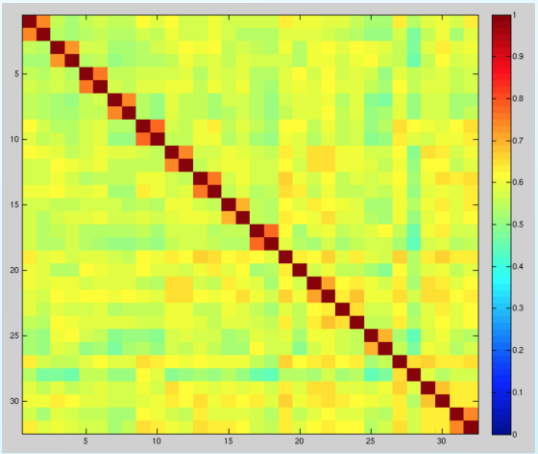
- After the first two data releases (6 months), 370 users have downloaded data. More than that have registered to browse through the data.
- A total of 41 terabytes (66,000 data files) have been downloaded.
- 38 laboratories have ordered (and paid for) the complete data set "Connectome in a Box"
- Multiple levels of processed data are available
  - Unprocessed (NIFTI)
  - Minimally preprocessed
  - Group-average Task-fMRI, functional connectivity
- After a very short period of time, outside users are reporting results using the Connectome data



# Human Connectome – Early Results




- For the first time, it appears that the imaging data from the HCP allows the brains from identical and/or fraternal twins to be distinguished from non-related individuals – this hasn't been achieved before.



National Institute of Mental Health

# Human Connectome Project




National Institute of Mental Health

# National Database For Autism Research

- NDAR is trying to solve a harder problem than the Human Connectome Project.
- The data in NDAR come from multiple laboratories and are measured for different purposes.
- There are many data collection instruments that measure similar clinical criteria.
- There are a number of cases where the same individual is seen in multiple different laboratories.
- There are other significant databases with autism data.

# National Database for Autism Research

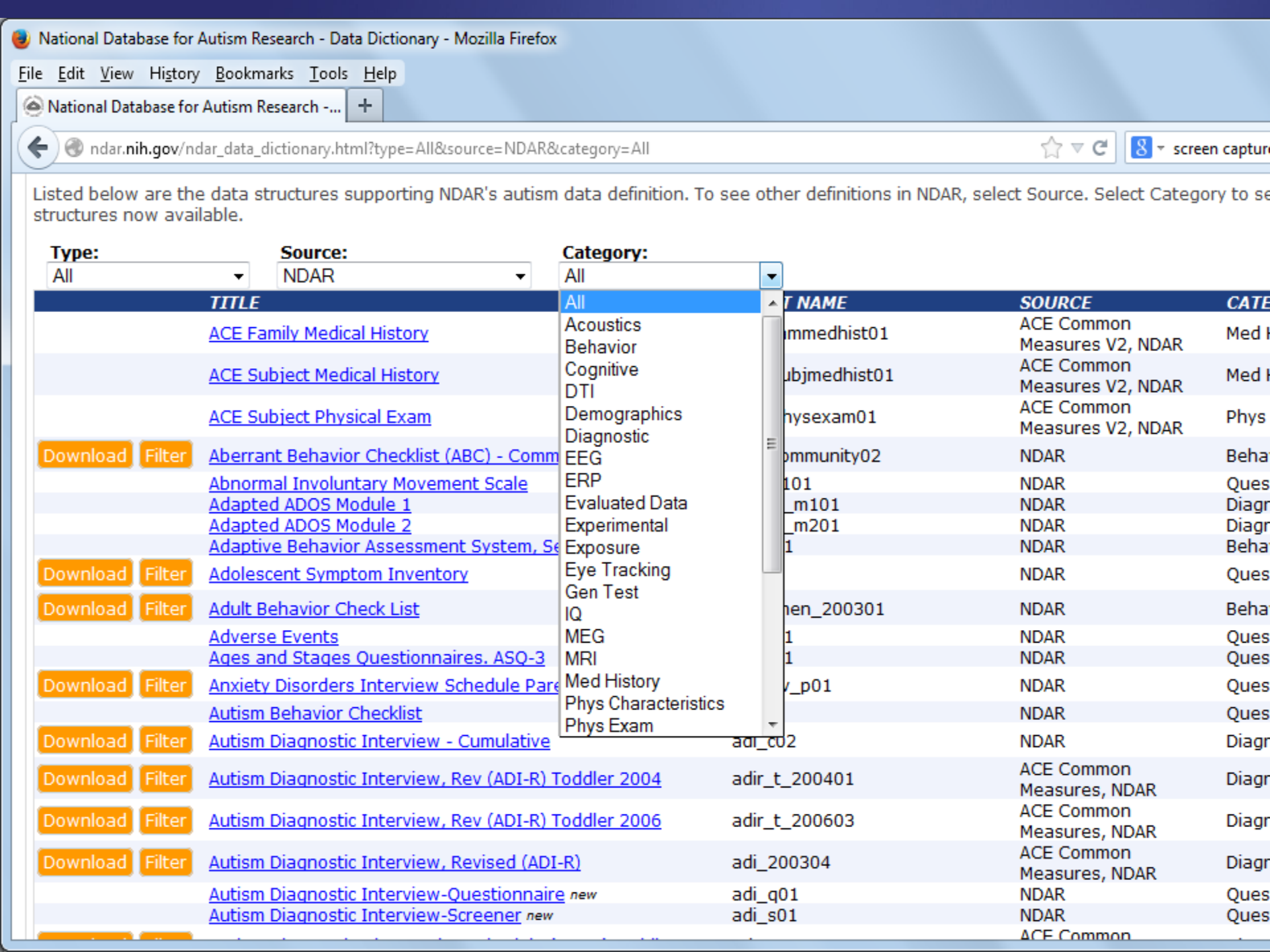
- Joint initiative supported by NIMH, NICHD, NINDS, and NIEHS
  - Federal data repository
  - Contains data from human subjects related to autism (and control subjects)
  - Data are available to the research community through a not too difficult application process
  - Summary data are available to everyone with a browser
- Begun in late 2006, first data was received in 2008, significant data became available in 2012.
- The data types include demographic data, clinical assessments, imaging data, eye tracking and other event related data, and –omic data
- Currently has data available from over 60,000 subjects
- 200+TB of imaging and –omic data is stored in the cloud

# NDAR: Implementation

- NDAR has deep federation with the following data repositories. This federation allows NDAR to query data in those repositories and to return data to the user from multiple repositories simultaneously.
  - Autism Tissue Program
  - Autism Genetic Resource Exchange
  - Interactive Autism Network
  - Simons Foundation Autism Research Initiative
- NDAR has two key features to allow data standardization and aggregation: data dictionaries and the Global Unique Identifier (GUID)
- Generally, NIH funded investigators are expected to share their data via NDAR. Investigators with funding from other sources are welcome to deposit their data.
- Over 80 studies have registered data, and more than 120 are expected to share data.

# Data Dictionary

- The NDAR data dictionary is one of the key building blocks for this repository. It provides a flexible and extensible framework for data definition by the research community.
- 400+ instruments, freely available to anyone
  - 50,000+ unique data elements and growing
  - A research community platform for defining the complex language characterizing autism research
    - Clinical
    - Genomics/Proteomics
    - Imaging Modalities
- Accommodates any data type and data structure
- Extended and enhanced by the ASD research community
- Curated by NDAR
- Allows investigators to quickly perform quality control tests of their data without submitting data anywhere.


National Database for Autism Research - Data Dictionary - Mozilla Firefox
File Edit View History Bookmarks Tools Help
National Database for Autism Research -...
ndar.nih.gov/ndar_data_dictionary.html?type=All&source=NDAR&category=All
Listed below are the data structures supporting NDAR's autism data definition. To see other definitions in NDAR, select Source. Select Category to se
structures now available.
Type:
All
Source:
NDAR
Category:
All
All
Acoustics
Behavior
Cognitive
DTI
Demographics
Diagnostic
EEG
ERP
Evaluated Data
Experimental
Exposure
Eye Tracking
Gen Test
IQ
MEG
MRI
Med History
Phys Characteristics
Phys Exam
TITLE
T NAME
SOURCE
CATE
ACE Family Medical History
mmedhist01
ACE Common Measures V2, NDAR
Med H
ACE Subject Medical History
ubjmedhist01
ACE Common Measures V2, NDAR
Med H
ACE Subject Physical Exam
hysexam01
ACE Common Measures V2, NDAR
Phys
Download
Filter
Aberrant Behavior Checklist (ABC) - Comm
mmunity02
NDAR
Beha
Abnormal Involuntary Movement Scale
101
NDAR
Ques
Adapted ADOS Module 1
_m101
NDAR
Diagn
Adapted ADOS Module 2
_m201
NDAR
Diagn
Adaptive Behavior Assessment System, Se
1
NDAR
Beha
Download
Filter
Adolescent Symptom Inventory
NDAR
Ques
Download
Filter
Adult Behavior Check List
hen_200301
NDAR
Beha
Adverse Events
1
NDAR
Ques
Ages and Stages Questionnaires. ASQ-3
1
NDAR
Ques
Download
Filter
Anxiety Disorders Interview Schedule Pare
_p01
NDAR
Ques
Autism Behavior Checklist
NDAR
Ques
Download
Filter
Autism Diagnostic Interview - Cumulative
adi_c02
NDAR
Diagn
Download
Filter
Autism Diagnostic Interview, Rev (ADI-R) Toddler 2004
adir_t_200401
ACE Common Measures, NDAR
Diagn
Download
Filter
Autism Diagnostic Interview, Rev (ADI-R) Toddler 2006
adir_t_200603
ACE Common Measures, NDAR
Diagn
Download
Filter
Autism Diagnostic Interview, Revised (ADI-R)
adi_200304
ACE Common Measures, NDAR
Diagn
Autism Diagnostic Interview-Questionnaire new
adi_q01
NDAR
Ques
Autism Diagnostic Interview-Screener new
adi_s01
NDAR
Ques

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| inter_med5 | Filter | String | 255 | Recommended | Current Medications 5 and dosage schedule | | | AB |
| ques_active | Filter | Integer | | Recommended | 1. Excessively active at home, school, work, or elsewhere | 0;1;2;3 | 0 = Not at all a problem; 1 = The behavior is a problem but slight in degree; 2 = The problem is moderately serious; 3 = The problem is severe in degree | AB |
| ques_selfinj | Filter | Integer | | Recommended | 2. Injures self on purpose | 0;1;2;3 | 0 = Not at all a problem; 1 = The behavior is a problem but slight in degree; 2 = The problem is moderately serious; 3 = The problem is severe in degree | AB |
| ques_inact | Filter | Integer | | Recommended | 3. Listless, sluggish, inactive | 0;1;2;3 | 0 = Not at all a problem; 1 = The behavior is a problem but slight in degree; 2 = The problem is moderately serious; 3 = The problem is severe in degree | AB |
| ques_aggres | Filter | Integer | | Recommended | 4. Aggressive to other children.or adults (verbally or physically) | 0;1;2;3 | 0 = Not at all a problem; 1 = The behavior is a problem but slight in degree; 2 = The problem is moderately serious; | AB |

National Database for Autism Research - Data Dictionary - Mozilla Firefox

File Edit View History Bookmarks Tools Help

National Database for Autism Research -...

ndar.nih.gov/ndar_data_dictionary.html?short_name=abc_community02

screen capture win

inter_med5

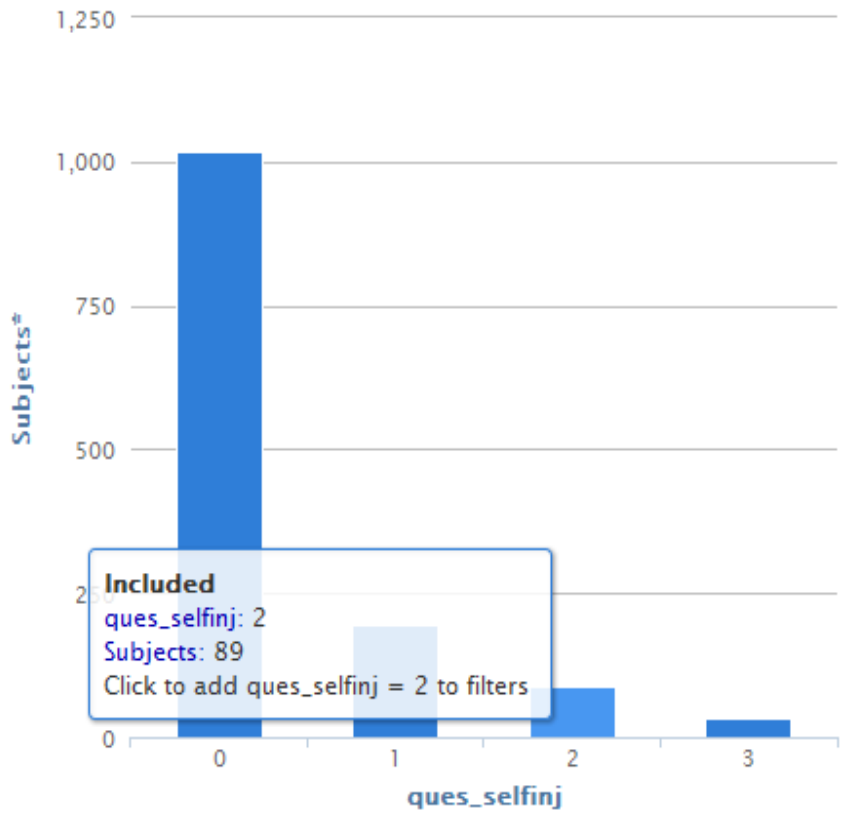
ques_active

ques_selfinj

ques_inact

ques_aggres

## Distribution for Data Structure: abc_community02 and Element: ques_selfinj




**Description**
2. Injures self on purpose

**Value Range**
0;1;2;3

**Notes**
0 = Not at all a problem; 1 = The behavior is a prob
slight in degree; 2 = The problem is moderately seri
The problem is severe in degree

*133 subjects have no value provided for ques_selfinj

**Filters**
No filters currently applied. Click bar on chart to add
filter.

Download R

# Global Unique Identifier

- The NDAR GUID software allows any researcher to generate a unique identifier using some information from a birth certificate.
- If the same information is entered in different laboratories, the same GUID will be generated.
- This strategy allows NDAR to aggregate data on the same subject collected in multiple laboratories without holding any of the personally identifiable information about that subject.
- The GUID is now being used in other research communities and can be made available to you.

# NDAR – Data Access

- Like the Human Connectome Project, researchers who obtain NIH funding for autism research are required to use consents that permit broad data sharing.
- NDAR receives written assurance that appropriate consents have been used by requiring both the PI and an institutional official to certify that this is true.
- Researchers (and anyone else) can browse the website without any limitation.
- In order to download data, a researcher from an institution with a FWA must submit a short request for access to NDAR. That request requires a signature from both the PI and an institutional official. The institution is required to "treat the data with respect."
- Data access is limited to a 1 year period. After that, a researcher needs to reapply.

# NDAR: Effect of Restricted Data Access

- Currently, 270 researchers have access to NDAR.
- Papers are being published using the data from NDAR, and grants are being funded to use that data.
- Rather then download data, NDAR encourages researchers to leave the data in the cloud and compute on it there. Currently, we have projects with multiple research groups to create pipelines to analyze MRI and genomics data in the cloud. Those pipelines will be freely available to the research community. It will be possible for researchers to combine their “private” data with data from NDAR in these pipelines.

# Conclusions

- It is possible, but not easy, to make information from human subjects research broadly available to the research community.
- Broad consents are the key to making human subjects data available.
- When data is collected by a researcher and then made available, a great deal of thought needs to go into who can access the data.
- If a “research subject” makes the data available directly (Patients Like Me, facebook, …), many of the concerns about data access disappear.